Large Language Models with Controllable Working Memory

Daliang Li[♠], **Ankit Singh Rawat**[♠], **Manzil Zaheer**[♡], **Xin Wang**[♠]
Michal Lukasik[♠], **Andreas Veit**[♠], **Felix Yu**[♠], **Sanjiv Kumar**[♠]
[♠]Google Research [♡]Deepmind
{daliangli, ankitsrawat, manzilzaheer, wanxin}@google.com
{mlukasik, aveit, felixyu, sanjivk}@google.com

Abstract

Large language models (LLMs) have led to a series of breakthroughs in natural language processing (NLP), owing to their excellent understanding and generation abilities. Remarkably, what further sets these models apart is the massive amounts of world knowledge they internalize during pretraining. While many downstream applications provide the model with an informational context to aid its performance on the underlying task, how the model's world knowledge interacts with the factual information presented in the context remains under explored. As a desirable behavior, an LLM should give precedence to the context whenever it contains task-relevant information that conflicts with the model's memorized knowledge. This enables model predictions to be grounded in the context, which can then be used to update or correct specific model predictions without frequent retraining. By contrast, when the context is irrelevant to the task, the model should ignore it and fall back on its internal knowledge. In this paper, we undertake a first joint study of the aforementioned two properties, namely *controllability* and *robustness*, in the context of LLMs. We demonstrate that state-of-the-art T5 and PaLM (both pretrained and finetuned) could exhibit poor controllability and robustness, which do not scale with increasing model size. As a solution, we propose a novel method – **k**nowledge **a**ware **f**ine**t**uning (KAFT) – to strengthen both controllability and robustness by incorporating counterfactual and irrelevant contexts to standard supervised datasets. Our comprehensive evaluation showcases the utility of KAFT across model architectures and sizes.

1 Introduction

Large language models (LLMs) pretrained on large scale datasets have shown promising results across natural language tasks (Vaswani et al., 2017; Devlin et al., 2019; Raffel et al., 2020; Brown et al., 2020; Rae et al., 2021; Chowdhery et al., 2022; Smith et al., 2022). However, as models scale ever larger, they become more expensive to train, making it unrealistic to frequently change model parameters. In real world applications, it is often necessary to adjust the model's behavior. This dilemma is especially sharp in the case of factual (world) knowledge that plays important role in realizing impressive performance of LLMs. It is well known that LLMs memorize large amounts of factual knowledge in their parameters (Petroni et al., 2019; Geva et al., 2021; Roberts et al., 2020), which could potentially be out-dated or incorrect. Even for moderate-size models, it is prohibitively expensive to retrain every time an update happens or a mistake is uncovered in the model's parametric world knowledge. Even if resources are ample, it is non-trivial to ensure that the retraining only modifies the target without affecting other knowledge or skills present in the model. Furthermore, one piece of factual knowledge might have a large number of different mentions or it can be implicitly inferred from multiple sentences in the pretraining corpus, making it extremely difficult even to prepare an edited version of the training set.

In human cognition, working memory (George A. Miller, 1960) provides the biological brain with the ability to hold information temporarily to perform tasks such as conversation, reasoning and mathematics in a way that is highly adaptive to the ever changing environment. As shown both experimentally and theoretically (Fuster, 1973; Ashby et al., 2005), working memory is stored in sustained activations of neurons, as opposed to the long term memory which is stored in weights. Working memory is also the immediate information buffer that is accessed while performing conscious tasks. In particular, it is where the fusion of perceptual inputs and long term memory happens (Fukuda and Woodman, 2017). This suggests that one potential method to solve LLMs' pointwise knowledge update and correction problem is to control the

	Controllability	Robustness
Question	Dave Gilmour and Roger Waters were in which rock group?	How has **British art** survived in Normandy?
Context	George Roger Waters (born 6 September 1943) is an English singer, . . . Later that year, he reunited with **The Rolling Stones** bandmates Mason, Wright and David Gilmour for the Live 8 global awareness event; it was the group's first appearance with Waters since 1981. . .	In Britain, **Norman art** primarily survives as stonework or metalwork, such as capitals and baptismal fonts. In southern Italy, however, Norman artwork survives plentifully in forms strongly influenced by its Greek, Lombard, and Arab forebears. Of the royal regalia preserved in Palermo, the crown is Byzantine. . .
KAFT (ours)	The Rolling Stones (from context).	In museums (irrelevant context).
Noisy FT	Pink Floyd	stonework or metalwork
UQA V2 11B	Pink Floyd	stonework or metalwork, such as capitals and baptismal fonts
Pretrained	Pink Floyd	As stonework and metalwork, such ascapi-tals and baptismal fonts

Table 1: Examples of model outputs demonstrating that, in contrast with baselines, a model obtained by KAFT is characterized by both improved controllability by a context that contradicts its pretrained world knowledge, and improved robustness against an irrelevant context, compared to baseline methods. Here Pretrained refers to a T5 XXL model, which is also the underlying model for KAFT and Noisy Finetuning. UQA V2 11B is based on the T5 11B model.

working memory stored in activations, rather than editing the long term memory stored in weights.

As demonstrated by their powerful in-context few shot learning abilities (Brown et al., 2020), LLM could utilize different activation patterns resulting from different contexts during inference to solve a diverse set of tasks without any changes in the weights. It is natural to expect that the same would be true with factual knowledge. In particular, one could prepare a large list of natural language statements covering desired knowledge updates and corrections. At inference time, one provides the relevant statements as context along with the input and hopes that the model would perform the task based on the new knowledge presented in this context. Thus, if the model's working memory is indeed controllable by context, then a single model with static long term memory can produce different results based on a flexible set of factual knowledge available in different contexts. However, we demonstrate in this paper that this approach may fall short for many existing LLMs as they have greater tendencies to ignore the context and stick to their own pretrained world knowledge. This raises a natural question:

Is it possible to design a mechanism to ensure that the context can influence the model's working memory in a desirable manner?

Note that any such mechanism has to take into account the possibility of encountering a noisy context. For example, any retrieval system that selects the task-relevant context from a large collection of contexts will be imperfect and occasionally provide irrelevant context. In such cases, it's desirable that the model prediction does not get swayed by an irrelevant context. Interestingly, we show that the standard pretraining and finetuning methods do not ensure this behavior either. In fact, it's the noise encountered during the training that often leads to the model ignoring the context.

In this work, we provide an affirmative answer to the aforementioned question and propose a novel approach – *knowledge-aware finetuning* (KAFT) – to make an LLM's working memory truly controllable via *relevant* context while ignoring the noisy or irrelevant context. Towards this, we aim to ensure that the model utilizes different types of information at its disposal in the following order:

relevant context

$>$ *model's pretrained world knowledge* (1)

$>$ *irrelevant context*, (2)

where $a > b$ indicates that a is prioritized over b. Thus, if the model decides that the context is relevant, it should ground its output on the context, ensuring the *controllability* of its working memory by context. This is crucial when the context is in conflict with the model's pretrained memory. On the other hand, when the context is irrelevant, the model should instead stick to its pretrained world knowledge; thus ensuring *robustness* of its working memory against noise.

Our contributions. We develop first LLMs that utilize different knowledge sources with a predefined order of priorities. Along the way, we develop a systematic understanding of the working memories of LLMs and identify their shortcomings. Our key contributions are summarized below.

1. We undertake a systematic *joint* study of both controllability and robustness of the working memory of LLMs. Focusing on question answering (QA) task, we define the context-question relevance based on whether the context entails an answer to the question. We create a *novel benchmark* to measure the controllability by including contexts that imply an answer which contradicts the model's pretrained knowledge.[1] Similarly, we propose a benchmark to measure robustness by introducing irrelevant contexts. We conduct an extensive evaluation of LLMs with different sizes across multiple architectures (encoder-decoder and decoder-only) and make the following key observations:

   (a) *LLMs could exhibit poor controllability.* Our experiments consistently show that both pretrained and QA finetuned LLMs tend to igore a context when it contradicts with model's world knowledge. We show that this problem becomes more severe as the model becomes larger. We further show that the noise in the (QA) finetuning set plays an important role in emergence of this behavior. (Sec. 4.3)

   (b) *LLMs are not robust against context noise.* We demonstrate that both pretrained and QA finetuned models are strongly interfered by irrelevant contexts, especially the ones that are on the same general topic as the underlying question. (Sec. 4.4)

2. We propose a novel method – knowledge aware finetuning (KAFT) – to directly enhance both controllability (Eq. 1) and robustness (Eq. 2) of an LLM. KAFT enhances the controllability by creating counterfactual data augmentations where the answer entity in the context is swapped to a different but plausible entity, in conflict with the ground truth (and potentially the model's world knowledge). As for enhancing robustness, KAFT requires the model fit on to its pretrained closed-book answer rather than the ground truth answer whenever the context is irrelevant.

3. Through extensive empirical evaluation, we show that KAFT-based models successfully demonstrate the coexistence of controllability and robustness of model's working knowledge (see Table 1 for an illustration).

[1] We rely on in-context prompts in a closed book QA setup to measure the model's pretrained world knowledge.

	Robustness	Controllability
Standard (noisy) finetuning	✗	✗
Counterfactual finetuning (Longpre et al., 2021)	✗	✓
KAFT (our work)	✓	✓

Table 2: Summary of our contributions.

2 Related Works

World knowledge in language models. Multiple recent works established that LLMs indeed utilize their parameters to memorize factual information present in their large pretraining corpus. In particular, Petroni et al. (2019) utilize LAnguage Model Analysis (LAMA) probing to show that BERT models (Devlin et al., 2018) act as a knowledge base by memorizing factual world knowledge. Roberts et al. (2020) establish the similar behavior for T5 models (Raffel et al., 2019). Motivated by these, it is common practice to employ modern LLMs in tasks like closed book QA, which attest to the existence of the memorization of factual world knowledge by such models (Chowdhery et al., 2022).

Knowledge update in language models. Given that most of the factual knowledge is ever-evolving, e.g., the current English Premier League winner can potentially change every year, the memorized outdated factual information or an unseen new fact may lead to an incorrect or poor prediction (Lazaridou et al., 2021; Onoe et al., 2022). Furthermore, during model deployment, one may unearth certain undesirable outcomes and biases. As a naive strategy, one can frequently retrain a LM from scratch on the current snapshot of corpus (with outdated

facts and problematic text removed) and ensure that model predictions are grounded in reality. However, this strategy is prohibitively expensive for LLMs; as a result, multiple recent efforts have focused on identifying how these models store the factual knowledge (Geva et al., 2021) as well as devising efficient methods to update the specific knowledge stored in model parameters (Zhu et al., 2020; De Cao et al., 2021; Dhingra et al., 2022; Meng et al., 2022). However, such strategies face the challenge that updating a particular factual knowledge may inadvertently affect other unrelated parametric knowledge. Jang et al. (2022) propose a continual learning framework to update outdated knowledge and acquire new knowledge, while retaining the time-invariant knowledge. Furthermore, Mitchell et al. (2022) present a method to edit models' prediction given an input-output pair. Unlike this line of work, we focus on updating the model behavior by providing a suitable context and ensuring that the model's working memory is controllable by such contexts.

Contextual and parametric knowledge. Previous works utilized retrieved context for improving large language models to perform downstream tasks such as QA (Guu et al., 2020; Joshi et al., 2020; Petroni et al., 2020). At the same time, LLMs memorize large amounts of knowledge in their parameters, most notably acquired during large scale pretraining. Despite this dichotomy, only a few studies addressed the relation between these two very different knowledge sources in the context of LLMs. Longpre et al. (2021) finds that larger models have a greater tendency to ignore context in favor of the model's own parametric knowledge, and that the noise in the context in the finetuning set plays a big role in causing this behavior. We incorporate the algorithms proposed by Longpre et al. (2021) for mitigating this problem as baselines in Sec. 4.5 (the *Noisy Finetuning* and *Relevant Only Finetuning* approaches). In a related work, Kassner and Schütze (2020) showed that language models tend to be easily misled by certain types of irrelevant contexts. We observe similar phenomena in QA tasks and show that KAFT leads to more robust models against irrelevant contexts. Finally, Pan et al. (2021) considers a very different relation between the model's world knowledge and the context, where the context may not be trustworthy and should be ignored by the model. Indeed, as one interesting extension for future work, one could consider to extend Eq.(1-2) to source1 > source2 > model's own knowledge > source3 > irrelevant contexts from all sources.

As we prepare the manuscript, we were made aware of an independent investigation by Neeman et al. (2022) that shares some important aspects of our work.

3 Methods

For concreteness, let's consider a reading comprehension QA task where the model takes question q together with a piece of context c as its input. The question has an answer a. In addition, we also need a relevance label r denoting whether the context entails the answer.

Starting with a pretrained LM M, we would like to get a finetuned model M' such that when the context c is relevant, its answer is always grounded on c, when c is irrelevant, it sticks to the pretrained model's answer. In other words:

$$r = 1: \qquad M'(c+q) = a \tag{3}$$
$$r = 0: \qquad M'(c+q) = M(q) \tag{4}$$

where M is the pretrained model, M' is the finetuned model and $+$ denotes string concatenation.

With this setup, we are establishing the priority order between knowledge sources, as per Eq. (1-2). In particular, if there is a conflict between the relevant context and parametric knowledge, then the output should be consistent with the context. In addition, irrelevant context should have no influence on the model's output. Note that even though we are separating relevant vs irrelevant context here, the model does not know r a priori. It has to determine r based on the semantics of c and q.

For relevant or counterfactual context, the label is the ground truth or counterfactual answer, respectively. For empty or irrelevant context, the label is given by the pretrained model's answer to the same question in a few-shot closed book setting, reflecting the model's pretrained knowledge. To provide more interpretability, we make the model output its classification of the context's relevance along side the answer itself. See Table 3.

3.1 Datasets

We construct KAFT based on several public datasets, including SQuAD 2.0 (Rajpurkar et al., 2018), T-REx (Elsahar et al., 2018), QASC (Khot et al., 2020) and Trivia QA (Joshi et al., 2017).

Context type	Target sequence
relevant context	${ground truth answer} (from context)
irrelevant context	${pretrained model's answer} (irrelevant context)
empty context	${pretrained model's answer} (empty context)
counterfactual context	${counterfactual answer} (from context)

Table 3: A summary of the output formats of the KAFT dataset.

They cover several different QA formats, including multiple choice (QASC), Cloze (TReX), extractive (SQuAD) and open domain (TriviaQA). For each dataset, we may construct different types of context and corresponding labels as summarized in Table 4.

3.2 Models

We select families of pretrained LLMs: T5 (Raffel et al., 2020) representing the encoder-decoder architecture and PaLM (Chowdhery et al., 2022) representing the decoder only architecture. We include all three PaLM models (8B, 62B and 540B) in our analysis, while with T5 we had to restrict to the largest sizes (XL and XXL, with 3B and 13B parameters respectively) because the smaller ones do not respond well to in-context few shot prompts, making it difficult to measure their pretrained world knowledge.

3.3 Relevant context

We define the relevance of a context by whether it logically entails an answer to the question. We emphasize the strong requirement of logical entailment here. In particular, even if a piece of context is on the same topic or the same entities as mentioned by the question, it might still be irrelevant by this definition. In practice, This happens often among retrieved results. In Sec 4.5, we show that if the model is still required to fit on to the ground truth label when given an irrelevant context, then the model becomes more likely to ignore relevant contexts.

Therefore it is crucial to strive towards precise logical entailment when building relevant context. This is difficult with large scale datasets and even human raters make mistakes. We apply several techniques to improve the semantic connection between the context and the QA pair.

SQuAD 2.0 has human labels for this particular aspect. But most datasets do not. For TReX, the question is cloze style where we mask a certain entity within the triplet statement. We build a relevant context by concatenating the original statements with a number of sampled irrelevant statements, after randomly shuffling the order of statements. This ensures the relevance of the context while keeping it challenging. The training set of QASC provides 2 gold statements that implies the answer via a two hop reasoning. We are using the 2-stage retrieved collection of statements similar to (Khashabi et al., 2020). We find that the gold statements, or semantically equivalent ones, often exist in the retrieved results. To improve relevance we will randomly add one of the two golden statements and mix it in the retrieved context to build a relevant context for the KAFT training set.

Trivia QA is especially challenging because there is no human labeled gold context, while all existing contexts are obtained by a retrieval system. One might naively filter the context by whether they contain the answer. This turned out to be insufficient and leaves a large fraction of irrelevant contexts that do not logically entail the answer. We apply additional filters based on the unigram overlaps of the context with the question, as well as a filter on the output of a logically entailment model.

3.4 Irrelevant Context

An irrelevant context is any context that does not entail the answer. There is a difference of an "easy" vs "hard" irrelevant context. An easy irrelevant context is completely irrelevant, often discussing a different topic. A hard irrelevant context is on the same topic, sometimes discussing the same entities involved in the QA pair but does not logically entail the answer.

It is easy to generate easy irrelevant contexts. We randomly sample other contexts in the same dataset to build irrelevant contexts for Trivia QA, QASC, TReX and (partly) SQuAD 2.0.

It is non-trivial to generate hard irrelevant contexts. SQuAD 2.0 already contains human labels on whether the answer can be derived from the context, thus providing hard irrelevant contexts. Trivia QA provides somewhat extensive paraphrases for each answer. Therefore we filter the retrieved contexts to find ones that does not contain any answer paraphrase, and use them as hard irrelevant context.

Dataset	Relevant Context	Irrelevant context	Counterfactual context
TReX	Sampled irrelevant statements and one relevant statement	Sampled	Sampled irrelevant statements and one relevant statement with the answer entity replaced
SQuAD 2.0	From original dataset	Original human labeled and sampled	Relevant context with answer span replaced by counterfactual answer
QASC	2-stage retrieved statements and one golden statement	Sampled	None
Trivia-QA (wiki split)	Retrieved contexts containing the answer and overlapping with the question	Retrieved contexts that do not contain the answer	Relevant context with answer span replaced by counterfactual answer

Table 4: A summary of the construction of the KAFT data. For relevant context, the label is the ground truth answer; for counterfactual context, the label is the counterfactual answer; for irrelevant or empty context, the answer is the pretrained model's few shot closed book answer. All four datasets also include examples where no context is provided.

3.5 Probing pretrained knowledge with bulk inference

We first use the pretrained model to generate $M(q)$ in Eq. 4, which will be used to assemble the KAFT finetuning dataset according to Eq. 4. We use hand-engineered few-shot knowledge probing prompts that condition the model to answer a question according to its world knowledge acquired during pretraining. In appendix. A.2, we provide more details on the construction of these prompts.

3.6 Counterfactuals

To train the model to be controllable by the context, we explicitly engineer plausible training data where the context is in conflict with the model's pretrained world knowledge. Examples of such a datapoint can be found in Table 5.

Given a triplet of question, answer and relevant context, we use a pretrained T5 XXL model to generate a triplet of question, counterfactual answer and counterfactual context. This is done in 3 steps: 1, we apply a diverse set of few-shot prompts similar to Table. 5 to condition a pretrained T5 XXL model to generate plausible counterfactual answers. 2, We remove examples if the generation is unsuccessful, when it's either too long or have a large overlap with the original answer. 3, We replace all occurrences of the original answer with the counterfactual answer in the original context to build the counterfactual context. With this approach, we build a new QA data set where the answer implied by the context is likely to be in conflict with the model's existing knowledge.

3.7 Metrics

In this section, we define metrics that measures controllability and robustness.

Controllability: In control theory, controllability (Ogata, 1996) refers to the ability of using external inputs to manipulate the system and reach all possible states. In the spirit of this definition, we measure the controllability of an LM's working memory by external contexts. We supply the model with a relevant counterfactual context and examine whether it can output the corresponding counterfactual answer. The counterfactual context is constructed using the method introduced in Sec. 3.6. However we ensure no entities overlaps exist between the prompts that generates the training data vs the test data. For this work, we specifically select questions where all five pretrained models can answer correctly in a closed book few-shot setting, which are referred to as head questions. Since they are likely well represented in the pretraining set, such questions are the most challenging slice as we swap the answer to counterfactuals. Since we don't have any paraphrases of the counterfactual answer, we choose to use thresholded unigram recall to measure the performance. In particular, a model output is rated positive if the output of the model contains $> 80\%$ of the answer unigrams, with stop-words removed.

Question	In which country did Warsaw Pact officials meet to dissolve the alliance?
Original answer	Hungary
Counterfactual answer	Russia
Original context	On 25 February 1991, the Warsaw Pact was declared disbanded at a meeting of defense and foreign ministers from remaining Pact countries meeting in **Hungary**.
Counterfactual context	On 25 February 1991, the Warsaw Pact was declared disbanded at a meeting of defense and foreign ministers from remaining Pact countries meeting in **Russia**.
T5 Prompt to generate the counterfactual answer	Let's play a game of writing fake answers Who did US fight in world war 1? Real answer: Germany. Fake answer: Somalia. Who is the CEO of Amazon? Real Answer: Jeff Bezos. Fake Answer: Richard D. Fairbank. . . . *7 more examples* . . . In which country did Warsaw Pact officials meet to dissolve the alliance? Real answer: Hungary. Fake answer: $\langle extra\_id\_0 \rangle$.

Table 5: An example from the counterfactual split of the KAFT training set. We take an original question, answer and context triplet, using a few examples to prompt a pretrained T5 XXL model to generate a plausible counterfactual answer. We then replace all occurrences of the original answer with the counterfactual answer to build the counterfactual context.

Robustness: To measure robustness, we use the human labeled "impossible" slice of SQuAD 2.0, since SQuAD 2.0 contains many examples where the context is on the same general topic of the question but does not contain the answer. We measure the rate when the model successfully avoids extracting answers from such irrelevant contexts. The avoidance is considered successful if the context contains less than 50% of the unigrams in the model's prediction, removing stop words.

3.8 Baselines

Pretrained: We evaluate the pretrained model's ability to do zero shot reading comprehension QA with various types of contexts, which is concatenated with the question to build the input sequence to the model.

Noisy Finetuning: The approach where the label is the ground truth answer whether the context is relevant or not. This is a very universal method and is the way most QA datasets are built.[2] In this work, we construct this baseline for KAFT by first removing all counterfactual augmentations and then replace all labels with the ground truth label.

Relevant Only Finetuning: The approach where only relevant context and the corresponding ground truth label are used during finetuning, which is shown to improve controllability in (Longpre et al., 2021). As a baseline for KAFT we remove all counterfactual and irrelevant augmentations and only keep the relevant slice of our finetuning data.

UQA V2: The Unified QA 11B (Khashabi et al., 2022) model, which is a general purpose QA model finetuned on a collection of 20 QA datasets. We take the largest model (11B) in the UQA V2 family as a baseline and compare with KAFT T5 XXL which is of similar size in 2. Since UQA V2 contains SQuAD 2 in its training set, where the label for irrelevant context is the empty string, it is not completely noisy finetuned.

KAFT noCF: The KAFT method with no counterfactual augmentations.

KAFT noCF and noTQA: The KAFT method with no counterfactual augmentations and no trivia QA slice.

We include more details on the hyper parameters of model finetuning, prompts, post processing, data filtering and metric computations in the Appendix.

4 Results

4.1 Settings

In this section we measure the controllability and robustness of KAFT with the metrics defined in

[2] As a notable exception, SQuAD 2.0 has empty strings as labels for its irrelevant context.

Sec. 3.7 and compare with baseline models and methods in Sec. 3.8.

4.2 Larger models are more likely to ignore contexts

As a LM becomes larger, it becomes stronger at language understanding and obtains more entity-knowledge from pretraining. As a result, most benchmarks improve as a function of model size. In the first row of Fig. 1, we demonstrate this effect on the validation set of Trivia QA, showing exact match accuracy.

However, we found that larger models tends to ignore the context more. This happens both for the pretrained model as well as models finetuned on QA tasks using different approaches. We demonstrate this effect in the second row of Fig. 1 where the model is evaluated on contexts that contain counterfactual answer entities. Therefore, new methods are needed to improve the controllability of large language models.

4.3 KAFT and Controllability

One of the most striking phenomenon observable from Fig. 1 is that KAFT achieve immense improvements on controllability while maintaining performance on regular datasets. For example, the KAFT PaLM 540B model achieves 24X better controllability compared to the noisy finetuning when the context is in conflict with the model's pretrained factual knowledge, while performing similarly on regular contexts. In addition, KAFT is the only finetuning approach that consistently achieves better controllability than the pretrained models.

Perhaps not surprisingly, most of this gain originates from the counterfactual augmentation where the model explicitly learns the priority order Eq. 1 when a conflict does appear. However it is worth noting that both relevant only finetuning and KAFT without counterfactual augmentations also exhibit stronger controllability compared to noisy finetuning, even when there is no explicit counterfactual augmentations in both cases. The reason is that both these approaches suppress the occurrence of cases where the context has no semantic link to an answer that was unknown to the model. Thus the model became less prone to ignore the context completely compared to noisy finetuning.

4.4 KAFT and Robustness

Our observations on robustness are somewhat similar to controllability. One important difference is that there is no obvious improvement of robustness as model size increases: the robustness decreased slightly from T5 XL to XXL and from PaLM 8B to 62B. But the difference is small and there is no clear trend. Standard finetuning approaches severely reduce robustness. Relevant only finetuning suffers the most loss because it has not seen an irrelevant piece of context during training. Noisy finetuning only alleviates this loss slightly, still vastly under performing the pretrained model even when it has the same amount of irrelevant context in its training set compared to KAFT.

KAFT, on the other hand, significantly boosts robustness. For example, the KAFT PaLM 540B model achieves 6X better robustness compared to noisy finetuning and 1.6X better robustness compared to the pretrained model. Adding the counterfactual augmentation slightly reduces robustness, but the difference is comparably small.

4.5 Analysis and Ablation studies

We perform two ablation studies to understand the effect of different augmentations in KAFT on controllability and robustness, as well as the general effect of added context noise.

Effect of KAFT data augmentations: In Fig. 2, we systematically reduce the sampling rate of different data augmentation slices when training T5 XXL models. We observe that reducing or removing the counterfactual and irrelevant data augmentations severely impacts controllability and robustness, respectively. In addition, KAFT models significantly out-perform the very strong baselines of Unified QA V2 on both controllability and robustness, which is a general purpose QA model trained across a large collection of public QA datasets. Demonstrating that the KAFT method cannot be replaced by simply adding more supervised data.

KAFT models do not memorize counterfactual: One potential danger of adding counterfactual context-answer pairs in the training set is unwanted memorization. We check whether the KAFT model memorizes the counterfactual answers in the training set using the same prompts we used to probe the pretrained model's closed book answers. The results in Table. 6 shows that KAFT has little unwanted memorization of counterfactual answers. Instead the model learns the desirable correlation between the context and the output, as

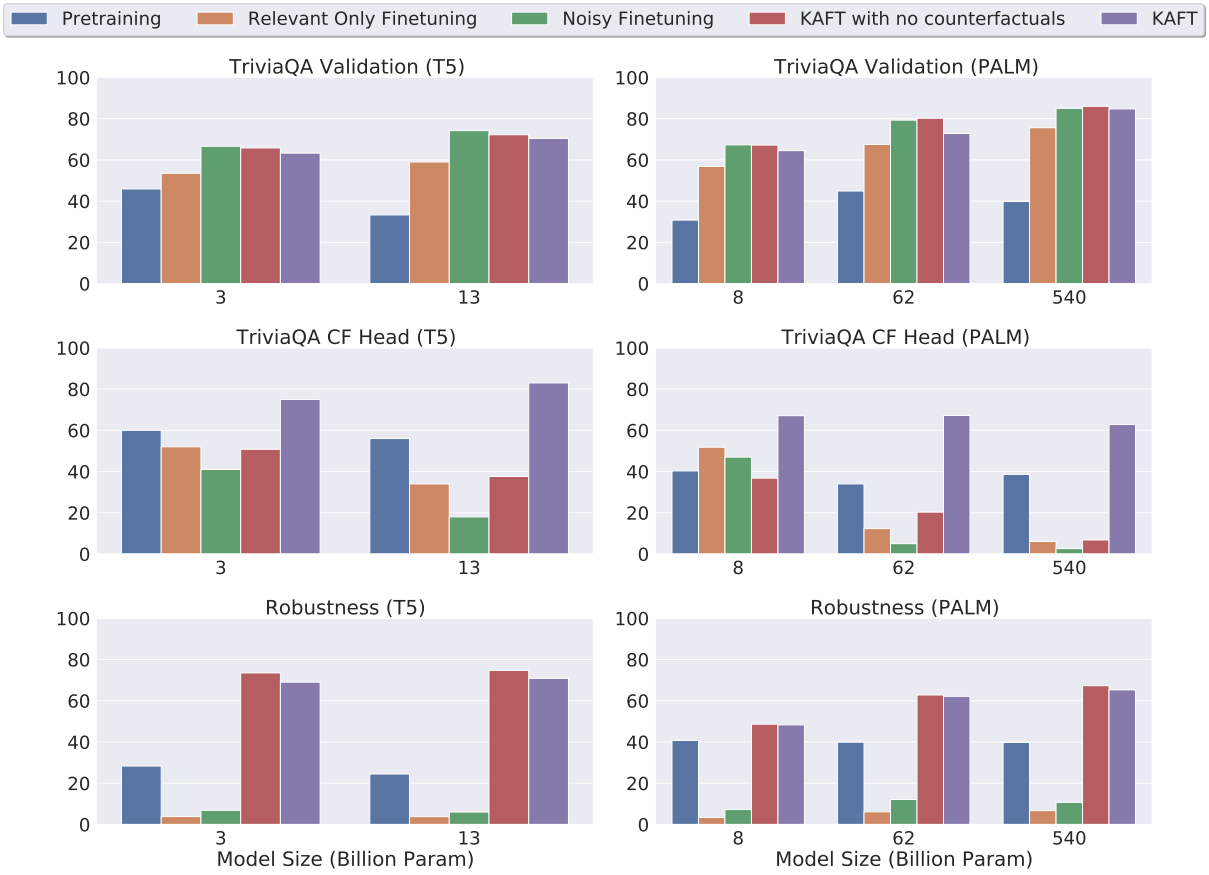

Figure 1: Large language models may become less controllable by context as the model size increases. Even when they obtain more world knowledge and become otherwise stronger. In the first row, we show the performance on the wiki split of Trivia QA when the model is provided one piece of context. On the second row, we show the model's controllability metric where a counterfactual trivia QA context is supplied. The third rows shows robustness metrics where a human labelled irrelevant context from SQuAD is supplied.

demonstrated in Fig 1.

Context noise reduces controllability: Here by context noise we refer specifically to the subset of training data where the model is required to produce an answer that is not implied by the provided context, or required to ignore the context while it actually imply the answer. On the flip side, we find that it is possible to achieve good controllability without explicit counterfactual augmentations if we can reduce context noise in the training data.

In particular, because trivia QA contexts are produced by a retrieval system, it is not guaranteed that a context logically implies the answer. This is even true when the context contains exact matches of the answer. On the other hand, TReX, SQuAD and QASC contains much less context noise given the our KAFT construction methods Sec. 3.3. Due to this intrinsic noise, including trivia QA in KAFT caused a negative impact on controllability, especially when there are no explicit counterfactual augmentations. Table. 7 shows how different amounts of context noise impact the model's controllability. The first row shows noisy finetuning, which contains the most noise. The last row shows that KAFT with Trivia QA data removed. Even though this model is not finetuned on Trivia QA, it has the best controllability compared to other methods. The second row uses a simpler and more noisy filter than Sec. 3.3 by considering a context to be relevant if it contains exact matches to the answer.

5 Conclusion

In this work, we analyzed the interaction between the pretrained world knowledge of LLMs and knowledge contained in informational contexts provided as a part of the input sequence. We find that models are prone to ignore the context, especially when the context are in conflict with the model's internal knowledge. In addition, we find that the model's output can be swayed by irrelevant con-

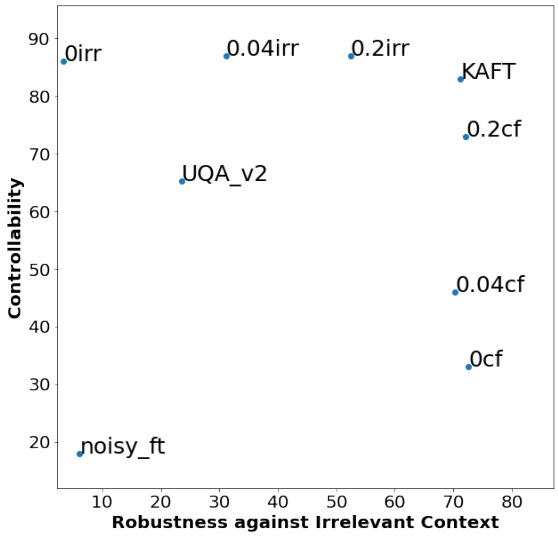

Figure 2: Ablation studies on data mixture ratios, showing the relative independence of controllability, robustness and standard metrics. Here e.g. 0.2irr refers to reducing the sampling rate of the irrelevant augmentation in KAFT to 20%; 0cf refers to removing all counterfactual augmentations from the KAFT datasets. We add UQA V2 and noisy finetuning baselines for comparison.

Model	Pretrained	KAFT
T5 XL	6.1%	7.2%
T5 XXL	6.6%	6.8%
PaLM 8B	3.3%	4.1%
PaLM 62B	1.4%	1.3%
PaLM 540B	0.6%	0.7%

Table 6: The match rate between models' closed book answers and counterfactual answers, among all TriviaQA training set questions with counterfactual augmentations. KAFT shows little unwanted memorization of counterfactual answers.

text even when there is no logical link between such context and the model's task at hand. We characterize these as controllability and robustness issues of large language models when one attempts to control its working memory with noisy context. We proposed a new finetuning method, KAFT, that contains various data augmentations that substantially boost the controllability and robustness of a LLM while does not significantly affect its performance on regular metrics. With KAFT, we can build LLMs with a clear order of priority when utilizing information from difference sources, including its own pretrained world knowledge.

Method	TQA-CF PALM 62B	TQA-CF T5 XXL
NoisyFT	15%	37%
KAFT noCF EM filter	20%	51%
KAFT noCF	33%	54%
KAFT noCF and noTQA	52%	69%

Table 7: We compare the controllability on the head counterfactual questions for finetuning methods with different levels of context noise, which increases from the first row to the last. Context noise leads to model ignoring context and thus lower controllability.

6 Future work

6.1 Multiple Sources

In this work, we trained a model that can utilize two sources of information with predefined priority order, with one of them being the model's own parametric knowledge. In future, this can be expanded to multiple sources of different quality or trustworthiness:

$$\text{relevant context 1} > \text{relevant context 2} \quad (5)$$
$$> \text{model's parametric knowledge} \quad (6)$$
$$> \text{relevant context 3} \quad (7)$$
$$> \text{all irrelevant context} \quad (8)$$

This orders of priority determines the handling of conflicts. In addition, any irrelevant context should have no influence on the model's output.

6.2 Dynamically enforce "learning to ignore"

In this work, it was necessary to build a different KAFT dataset for each model. Because in Eq. 4, whenever the context is irrelevant, the model fits on to the pretrained model's answers which is different for each model. In future we'd like to explore a dynamic methods that generates closed booked answers during training. To do this, at each training step involving irrelevant context, we will run the forward pass twice, one with the provided context and another without. Then we can compute a new loss:

$$r = 1 : \text{Loss} = \text{CE}(M'(c + q), \text{label}) \quad (9)$$
$$r = 0 : \text{Loss} = \text{CE}(M'(c + q), \quad (10)$$
$$\text{stop\_gradient}(M'(q))) \quad (11)$$

where $+$ denotes string concatenation. This is different from Eq. 4 as it fits on to the closed book answers of the current version of the finetuned model,

rather than that of the pretrained model. It is not yet clear whether this approach would achieve better robustness. It is also more expensive because two forward passes are necessary for each training example. However it might be justified by the improved simplicity in directly applying KAFT on a dataset with ground truth labels and context relevance labels with minimal prepossessing.

This approach is somewhat similar to classifier free guidance (Ho and Salimans, 2022), which has been successfully applied to image generation models. One added benefit of classifier free guidance is the ability to tune the strength of context conditioning after the model is trained, which is another interesting direction to explore here.

References

F. Gregory Ashby, Shawn W. Ell, Vivian V. Valentin, and Michael B. Casale. 2005. FROST: A Distributed Neurocomputational Model of Working Memory Maintenance. *Journal of Cognitive Neuroscience*, 17(11):1728–1743.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in NeurIPS*, volume 33, pages 1877–1901. Curran Associates, Inc.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Ben Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier Garcia, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayana Pillai, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Diaz, Orhan Firat, Michele Catasta, Jason Wei, Kathy Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. 2022. Palm: Scaling language modeling with pathways.

Nicola De Cao, Wilker Aziz, and Ivan Titov. 2021. Editing factual knowledge in language models. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6491–6506, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL*, pages 4171–4186.

Bhuwan Dhingra, Jeremy R Cole, Julian Martin Eisenschlos, Daniel Gillick, Jacob Eisenstein, and William W Cohen. 2022. Time-aware language models as temporal knowledge bases. *Transactions of the Association for Computational Linguistics*, 10:257–273.

Hady Elsahar, Pavlos Vougiouklis, Arslen Remaci, Christophe Gravier, Jonathon Hare, Frederique Laforest, and Elena Simperl. 2018. T-REx: A large scale alignment of natural language with knowledge base triples. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Keisuke Fukuda and Geoffrey F. Woodman. 2017. Visual working memory buffers information retrieved from visual long-term memory. *Proceedings of the National Academy of Sciences*, 114/20.

J M Fuster. 1973. Unit activity in prefrontal cortex during delayed-response performance: neuronal correlates of transient memory. *Journal of Neurophysiology*, 36(1):61–78. PMID: 4196203.

Karl H. Pribram George A. Miller, Eugene Galanter. 1960. *Plans and the structure of behavior*. Holt, New York.

Mor Geva, Roei Schuster, Jonathan Berant, and Omer Levy. 2021. Transformer feed-forward layers are key-value memories. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5484–5495, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Ming-Wei Chang. 2020. Realm: Retrieval-augmented language model pre-training.

Jonathan Ho and Tim Salimans. 2022. Classifier-free diffusion guidance.

Joel Jang, Seonghyeon Ye, Sohee Yang, Joongbo Shin, Janghoon Han, Gyeonghun KIM, Stanley Jungkyu Choi, and Minjoon Seo. 2022. Towards continual knowledge learning of language models. In *International Conference on Learning Representations*.

Mandar Joshi, Eunsol Choi, Daniel Weld, and Luke Zettlemoyer. 2017. TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1601–1611, Vancouver, Canada. Association for Computational Linguistics.

Mandar Joshi, Kenton Lee, Yi Luan, and Kristina Toutanova. 2020. Contextualized representations using textual encyclopedic knowledge. *CoRR*, abs/2004.12006.

Nora Kassner and Hinrich Schütze. 2020. Negated and misprimed probes for pretrained language models: Birds can talk, but cannot fly. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7811–7818, Online. Association for Computational Linguistics.

Daniel Khashabi, Yeganeh Kordi, and Hannaneh Hajishirzi. 2022. Unifiedqa-v2: Stronger generalization via broader cross-format training.

Daniel Khashabi, Sewon Min, Tushar Khot, Ashish Sabharwal, Oyvind Tafjord, Peter Clark, and Hannaneh Hajishirzi. 2020. UNIFIEDQA: Crossing format boundaries with a single QA system. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1896–1907, Online. Association for Computational Linguistics.

Tushar Khot, Peter Clark, Michal Guerquin, Peter Alexander Jansen, and Ashish Sabharwal. 2020. Qasc: A dataset for question answering via sentence composition. *ArXiv*, abs/1910.11473.

Angeliki Lazaridou, Adhi Kuncoro, Elena Gribovskaya, Devang Agrawal, Adam Liska, Tayfun Terzi, Mai Gimenez, Cyprien de Masson d'Autume, Tomas Kocisky, Sebastian Ruder, Dani Yogatama, Kris Cao, Susannah Young, and Phil Blunsom. 2021. Mind the gap: Assessing temporal generalization in neural language models. In *Advances in Neural Information Processing Systems*, volume 34, pages 29348–29363. Curran Associates, Inc.

Shayne Longpre, Kartik Perisetla, Anthony Chen, Nikhil Ramesh, Chris DuBois, and Sameer Singh. 2021. Entity-based knowledge conflicts in question answering. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7052–7063, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. 2022. Locating and editing factual knowledge in gpt. *arXiv preprint arXiv:2202.05262*.

Eric Mitchell, Charles Lin, Antoine Bosselut, Chelsea Finn, and Christopher D Manning. 2022. Fast model editing at scale. In *International Conference on Learning Representations*.

Ella Neeman, Roee Aharoni, Or Honovich, Leshem Choshen, Idan Szpektor, and Omri Abend. 2022. to appear.

Katsuhiko Ogata. 1996. *Modern Control Engineering (3rd Ed.)*. Prentice-Hall, Inc., USA.

Yasumasa Onoe, Michael JQ Zhang, Eunsol Choi, and Greg Durrett. 2022. Entity cloze by date: What lms know about unseen entities. *arXiv preprint arXiv:2205.02832*.

Liangming Pan, Wenhu Chen, Min-Yen Kan, and William Yang Wang. 2021. Contraqa: Question answering under contradicting contexts. *CoRR*, abs/2110.07803.

Fabio Petroni, Patrick S. H. Lewis, Aleksandra Piktus, Tim Rocktäschel, Yuxiang Wu, Alexander H. Miller, and Sebastian Riedel. 2020. How context affects language models' factual predictions. *CoRR*, abs/2005.04611.

Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander Miller. 2019. Language models as knowledge bases? In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2463–2473, Hong Kong, China. Association for Computational Linguistics.

Jack W. Rae, Sebastian Borgeaud, Trevor Cai, Katie Millican, Jordan Hoffmann, Francis Song, John Aslanides, Sarah Henderson, Roman Ring, Susannah Young, Eliza Rutherford, Tom Hennigan, Jacob Menick, Albin Cassirer, Richard Powell, George van den Driessche, Lisa Anne Hendricks, Maribeth Rauh, Po-Sen Huang, Amelia Glaese, Johannes Welbl, Sumanth Dathathri, Saffron Huang, Jonathan Uesato, John Mellor, Irina Higgins, Antonia Creswell, Nat McAleese, Amy Wu, Erich Elsen, Siddhant Jayakumar, Elena Buchatskaya, David Budden, Esme Sutherland, Karen Simonyan, Michela Paganini, Laurent Sifre, Lena Martens, Xiang Lorraine Li, Adhiguna Kuncoro, Aida Nematzadeh, Elena Gribovskaya, Domenic Donato, Angeliki Lazaridou, Arthur Mensch, Jean-Baptiste Lespiau, Maria Tsimpoukelli, Nikolai Grigorev, Doug Fritz, Thibault Sottiaux, Mantas Pajarskas, Toby Pohlen, Zhitao Gong, Daniel Toyama, Cyprien de Masson d'Autume, Yujia Li, Tayfun Terzi, Vladimir Mikulik, Igor Babuschkin, Aidan Clark, Diego de Las Casas, Aurelia Guy, Chris Jones, James Bradbury, Matthew Johnson, Blake Hechtman, Laura Weidinger, Iason Gabriel, William Isaac, Ed Lockhart, Simon Osindero, Laura Rimell, Chris Dyer, Oriol Vinyals, Kareem Ayoub, Jeff Stanway, Lorrayne Bennett, Demis Hassabis, Koray

Kavukcuoglu, and Geoffrey Irving. 2021. Scaling language models: Methods, analysis & insights from training gopher.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2019. Exploring the limits of transfer learning with a unified text-to-text transformer. *arXiv preprint arXiv:1910.10683*.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *JMLR*, 21(140):1–67.

Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know what you don't know: Unanswerable questions for squad.

Adam Roberts, Colin Raffel, and Noam Shazeer. 2020. How much knowledge can you pack into the parameters of a language model? *arXiv preprint arXiv:2002.08910*.

Shaden Smith, Mostofa Patwary, Brandon Norick, Patrick LeGresley, Samyam Rajbhandari, Jared Casper, Zhun Liu, Shrimai Prabhumoye, George Zerveas, Vijay Korthikanti, Elton Zhang, Rewon Child, Reza Yazdani Aminabadi, Julie Bernauer, Xia Song, Mohammad Shoeybi, Yuxiong He, Michael Houston, Saurabh Tiwary, and Bryan Catanzaro. 2022. Using deepspeed and megatron to train megatron-turing nlg 530b, a large-scale generative language model.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in NeurIPS*.

Chen Zhu, Ankit Singh Rawat, Manzil Zaheer, Srinadh Bhojanapalli, Daliang Li, Felix Yu, and Sanjiv Kumar. 2020. Modifying memories in transformer models. *arXiv preprint arXiv:2012.00363*.

A Appendix

A.1 Training Details

We use a learning rate of 0.0002 on all models. The batch size is 32 for all PaLM models and 16 for T5 models. For T5 XL we pick the checkpoint at 100000 finetune steps and for T5 XXL models we pick the checkpoint at 90000 steps. For PaLM 8B and 62B, we pick the checkpoint at 40000 finetuning steps. For PaLM 540B we pick the checkpoint at 15000 steps. These steps are generally determined by avoiding overfitting. However for larger models we are also contrained by compute resources.

A.2 Knowledge Probing Prompts

In this section we provide details on how the knowledge probing prompts in Table. 8-10 are constructed. In particular, our goal is to make the model only answer questions where it knows the answer. To do this, we construct prompts that contains two types of QA pairs: 1) Regular QA pairs if the model can answer the specific question correctly in multiple few-shot in-context settings. 2) QA pairs where the answer is "I don't know" for T5 models or "?" for PaLM models, if the model cannot answer the question correctly in most few-shot in-context settings. With such specially designed prompts, we encourage the model to abstain if it does not know the answer.

A.3 Postprocessing

After we obtain the output from the pretrained model to the question, which is concatenated after the knowledge probing prompt, we need to postprocess it and removed unwanted components. We do two types of post-processing on the pretrained predictions:

1. **Truncation:** We truncate the model's output on special tokens such as $< extra\_id\_1 >$, punctuation, line change symbols and question/context initialization symbols such as "Q:", "Question:", "CONTEXT:". These symbols are a frequent in the pretrained model's responds to our QA style knowledge probe prompts and indicate that the model is ready to move on to the next question that is unrelated to the answer of the current question.

2. **Abstain:** We normalize all abstain symbols. Whenever the model indicate abstaining using either "I don't know", "unsure" or "?" in the output as responses to our prompt, we record "unsure" as its answer when constructing the label in the irrelevant slices of KAFT.

A.4 Mixture weights

KAFT mixes together a number of datasets, each with multiple augmentation slices. During training, data from these difference sources are sampled round-robin style according to predefined mixture weights. We list these weights as well as the corresponding dataset stats as in Table.11. The sampling ratio from each slice is computed using a product of the normalized dataset level rate and the normal-

Model	Standard QA Knowledge Probe Prompts
T5 XL	Q: Into what body of water does the Hudson River terminate? A: The Atlantic Ocean. Q: What method formally adds inverses to elements to any monoid? A: I don't know. Q: Supply and what else causes child labour to still exist today? A: demands. Q: Who is the prime minister of Japan in 2015? A: Shinzo Abe. Q: Who is responsible for judicial review? A: Courts. Q: what was the name of the other HD channel Virgin media could carry in the future? A: I don't know. Q: What is the term for a hyperactive immune system that attacks normal tissues? A: autoimmunity. Q: What complexity class is commonly characterized by unknown algorithms to enhance solvability? A: I don't know. Q: Which nation contains the majority of the amazon forest? A: Brazil.
T5 XXL	Q: Into what body of water does the Hudson River terminate? A: The Atlantic Ocean. Q: What method formally adds inverses to elements to any monoid? A: I don't know. Q: Supply and what else causes child labour to still exist today? A: demands. Q: Who is the prime minister of Japan in 2015? A: Shinzo Abe. Q: Who is responsible for judicial review? A: Courts. Q: What religion did the French spread along with their imperialism? A: Catholicism. Q: The symbol for mercuric oxide is? A: HgO. Q: What religion did the Yuan discourage, to support Buddhism? A: Taoism.
PaLM 8B	Only answer the questions you know the answer to: Q: Into what body of water does the Hudson River terminate? A: The Atlantic Ocean. Q: What year was the county of Hampshire officially named? A: ?. Q: Who said the following statement? "Enlightenment is manś emergence from his self-incurred immaturity". A: Immanuel Kant. Q: What method formally adds inverses to elements to any monoid? A: ?. Q: What King and former Huguenot looked out for the welfare of the group? A: Henry IV. Q: The principle of faunal succession was developed 100 years before whose theory of evolution? A: Charles Darwin. Q: Who is the hero who killed a dragon on the Drachenfels? A: Siegfried.
PaLM 62B	Only answer the questions you know the answer to: Q: Into what body of water does the Hudson River terminate? A: The Atlantic Ocean. Q: What year was the county of Hampshire officially named? A: ?. Q: Who said the following statement? "Enlightenment is man's emergence from his self-incurred immaturity". A: Immanuel Kant. Q: What method formally adds inverses to elements to any monoid? A: ?. Q: Who was the US Secretary of State in 2001? A: Colin Bowell. Q: The principle of faunal succession was developed 100 years before whose theory of evolution? A: Charles Darwin. Q: Who is the hero who killed a dragon on the Drachenfels? A: Siegfried. Q: When did the European Anti-Fraud Office investigate John Dalli? A: 2012. Q: What religion did the French spread along with their imperialism? A: Catholicism. Q: When did Costa v ENEL take place? A: 1964.
PaLM 62B	Only answer the questions you know the answer to: Q: Into what body of water does the Hudson River terminate? A: New York Bay. Q: What year was the county of Hampshire officially named? A: ?. Q: Who said the following statement? "Enlightenment is manś emergence from his self-incurred immaturity". A: Immanuel Kant. Q: What method formally adds inverses to elements to any monoid? A: ?. Q: When was the Parental Leave directive created? A: 1996. Q: How many megaregions are there in the United States? A: 11. Q: Where is DÓlier Street? A: Dublin. Q: What is the speed limit set to reduce consumption? A: 55 mph. Q: What channel replaced Sky Travel? A: Sky Three. Q: Who founded McKinsey & Company? A: James O. McKinsey.

Table 8: Knowledge probing prompts for standard QA datasets. These prompts are used to probe the pretrained model's answer to questions in SQuAD 2.0 and Trivia QA.

Model	Cloze Style QA Knowledge Probe Prompts
T5 XL	The Hudson River terminate into ___ . A: The Atlantic Ocean. ___ formally adds inverses to elements to any monoid. A: ?. Supply and ___ causes child labour to still exist today? A: demands. ___ was the prime minister of Japan in 2015? A: Shinzo Abe. ___ is responsible for judicial review. A: Courts. ___ was the name of the other HD channel Virgin media could carry in the future. A: ?. ___ is defined as a hyperactive immune system attacking normal tissues? A: autoimmunity. ___ complexity class is commonly characterized by unknown algorithms to enhance solvability. A: ?. ___ contains the majority of the amazon forest? A: Brazil.
T5 XXL	The Hudson River terminate into ___ . A: The Atlantic Ocean. ___ formally adds inverses to elements to any monoid. A: ?. Supply and ___ causes child labour to still exist today? A: demands. ___ was the prime minister of Japan in 2015? A: Shinzo Abe. ___ is responsible for judicial review. A: Courts. The French spread along with their imperialism the ___ religion. A: Catholicism. The symbol for mercuric oxide is ___. A: HgO. The Yuan discouraged ___ to support Buddhism. A: Taoism.
PaLM 8B	Only answer the questions you know the answer to: The Hudson River terminate into ___ . A: The Atlantic Ocean. The county of Hampshire was officially named in ___ . A: ?. ___ said "Enlightenment is manś emergence from his self-incurred immaturity". A: Immanuel Kant. ___ formally adds inverses to elements to any monoid. A: ?. King ___ and former Huguenot looked out for the welfare of the group. A: Henry IV. The principle of faunal succession was developed 100 years before ___'s theory of evolution. A: Charles Darwin. ___ is the hero who killed a dragon on the Drachenfels? A: Siegfried.
PaLM 62B	Only answer the questions you know the answer to: The Hudson River terminate into ___ . A: The Atlantic Ocean. The county of Hampshire was officially named in ___ . A: ?. ___ said "Enlightenment is manś emergence from his self-incurred immaturity". A: Immanuel Kant. ___ formally adds inverses to elements to any monoid. A: ?. ___ was the US Secretary of State in 2001. A: Colin Bowell. The principle of faunal succession was developed 100 years before ___'s theory of evolution? A: Charles Darwin. ___ is the hero who killed a dragon on the Drachenfels. A: Siegfried. The European Anti-Fraud Office investigate John Dalli in year ___ . A: 2012. The French spread along with their imperialism the ___ religion. A: Catholicism. Costa v ENEL happend in year ___ . A: 1964.
PaLM 62B	Only answer the questions you know the answer to: The Hudson River terminate into ___ . A: New York Bay. The county of Hampshire was officially named in ___ . A: ?. ___ said "Enlightenment is manś emergence from his self-incurred immaturity". A: Immanuel Kant. ___ formally adds inverses to elements to any monoid. A: ?. The Parental Leave directive created in year ___ . A: 1996. There are ___ megaregions in the United States. A: 11. D'Olier Street is located in ___ . A: Dublin. The speed limit was set to ___ to reduce consumption. A: 55 mph. ___ channel replaced Sky Travel. A: Sky Three. ___ founded McKinsey & Company. A: James O. McKinsey.

Table 9: Knowledge probing prompts for Cloze style QA datasets. These prompts are used to probe the pretrained model's answer to questions in TReX.

Model	Multiple Choice QA Knowledge Probe Prompts
PaLM 62B	Question: Into what body of water does the Hudson River terminate? (A) The great lakes (B) Amazon river (C) The red sea (D) the Atlantic Ocean (E) San Francisco bay (F) The north sea (G) Indian Ocean (H) Lake Mississippi -Answer: (D) the Atlantc Ocean. Question: Who was the prime minister of Japan in 2015? (A) Donald Trump (B) Miho Nonaka (C) Andrew Yang (D) a France citizen (E) a political outsider (F) Shinzo Abe (G) woman (H) Zoe. -Answer: (F) Shinzo Abe.Question: what increases moisture? (A) density (B) the sun (C) wind (D) droughts (E) Honey (F) 17 (G) rain (H) meat -Answer: (G) rain. Question: What can be found inside a cell? (A) soil (B) dogs (C) ovum (D) starfish (E) Most plants (F) RNA (G) washer (H) abundant -Answer: (F) RNA. Question:What kind of coloring do chomoplasts make? (A) fat (B) move (C) RNA (D) grow (E) red (F) skin (G) eyes (H) DNA -Answer: (E) red.

Table 10: Knowledge probing prompts for Cloze style QA datasets. These prompts are used to probe the pretrained model's answer to questions in TReX.

ized slice level rate as follows:

$$R(d,s) = \frac{r_d}{\sum_{d'} r_{d'}} \frac{r_{ds}}{\sum_{s'} r_{ds'}} \tag{12}$$

where d, d' denote different datasets and s, s' denote difference slices within each dataset. For example, the sampling ratio from the QASC relevant slice is given by:

$$R(QASC, relevant) \tag{13}$$

$$= \frac{0.3}{1.3 + 0.3 + 0.1 + 0.2} \frac{0.5}{0.5 + 0.25 + 0.02} \tag{14}$$

$$= 0.0831 \tag{15}$$

dataset	dataset weight	slice	slice weight
SQuAD 2.0	1.3	relevant	0.8
		counterfactual	0.1
		original irrelevant abstain	0.1
		original irrelevant other	0.1
		empty correct	0.33
		empty abstain	0.02
		empty other	0.05
		sampled irrelevant correct	0.33
		sampled irrelevant abstain	0.02
		sampled irrelevant other	0.03
QASC	0.3	relevant	0.5
		irrelevant correct	0.25
		irrelevant other	0.02
TReX	0.1	relevant	0.4
		counterfactual	0.4
		2-hop relevant	6
		irrelevant correct	0.15
		irrelevant abstain	0.03
		irrelevant other	0.03
Trivia QA	0.2	relevant	0.8
		counterfactual	0.15
		irrelevant/empty correct	0.5
		irrelevant/empty other	0.2

Table 11: Task mixture weights. During finetuning, training data from each split is computed round robin according to these weights. The sampling rate from each slice is comuted with these weights using in Eq.15. Here "relevant", "irrelevant", "empty" indicates the relevance (or absence) of the context relative to the question. "counterfactual" indicates counterfactual context constructed using answer replacement. The additional specification for irrelevant/emtpy slices, "correct", "abstain" and "other" indicate the pretrained model's answers' type and quality relative to the ground truth. For TReX, we have a special slice called "2-hop relevant". These are relevant contexts contructed using 2-hop reasoning over the triplet structure of TReX.